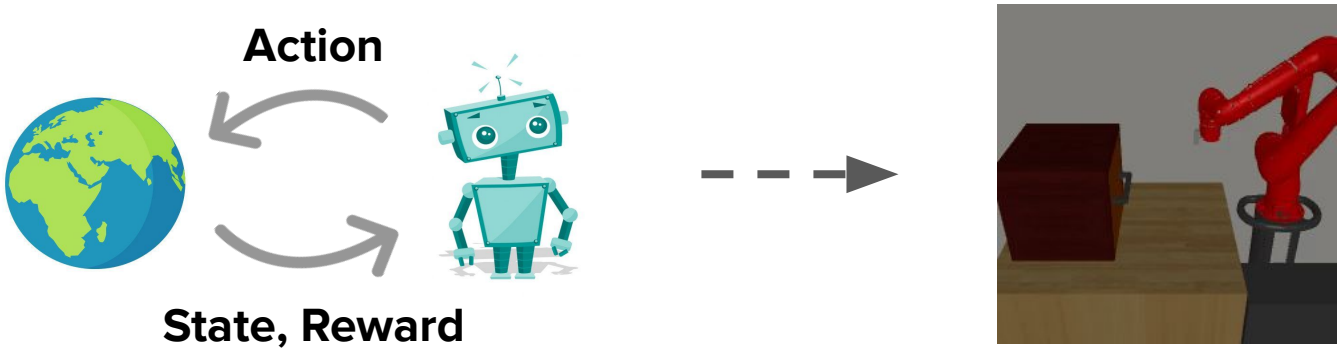
MURAL: Meta-Learning Uncertainty-Aware Rewards for Outcome-Driven Reinforcement Learning

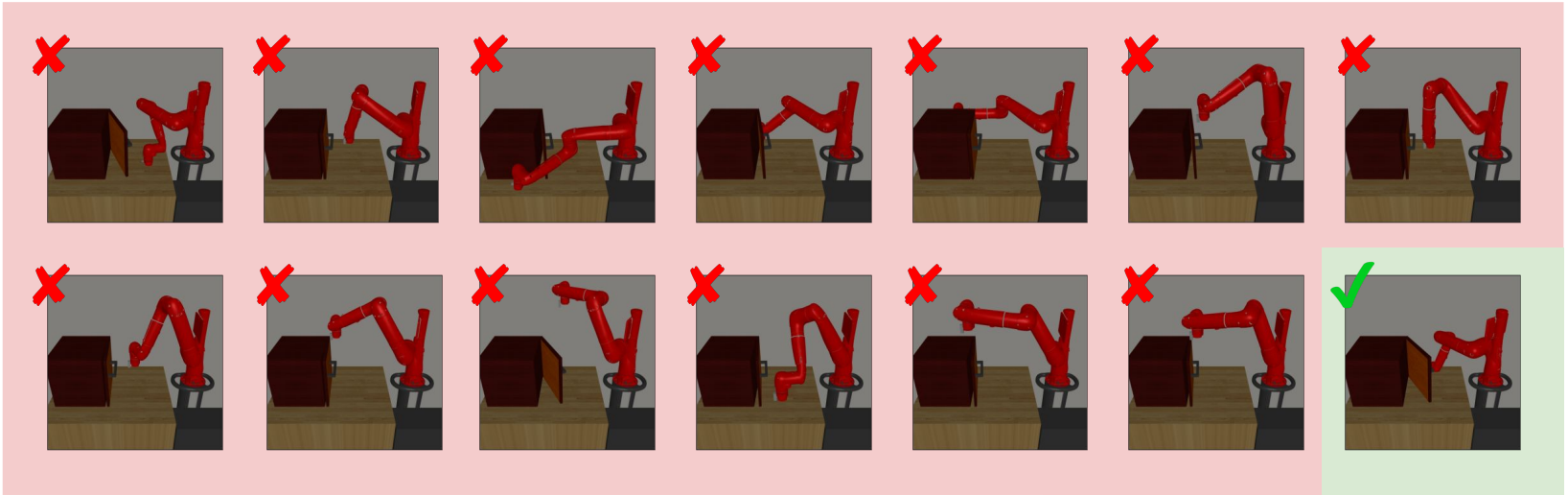
Kevin Li*, Abhishek Gupta*, Ashwin Reddy, Vitchyr Pong, Aurick Zhou, Justin Yu, Sergey Levine

* = equal contribution

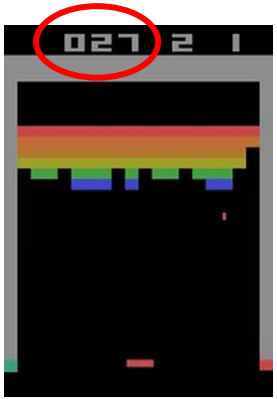
Reinforcement learning is a powerful framework for learning skills through interaction with the environment.



Often intractable — too many states to check!

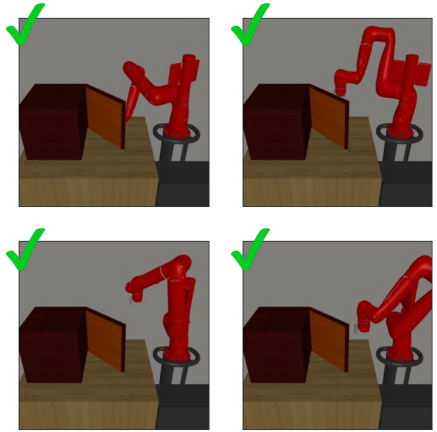


Reward shaping is helpful, but not always possible...



???

What if we formulate the RL problem in terms of **successful outcomes?**

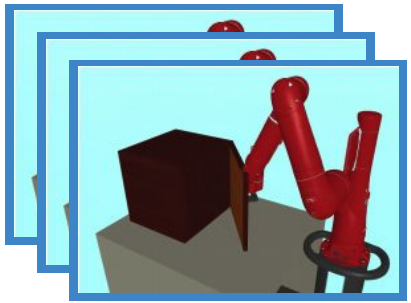


Success: door is open at 45 degrees

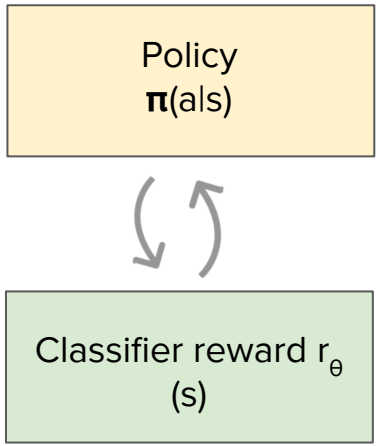
Classifier-Based Rewards

Idea: *provide a set of desired outcomes, then train a classifier to distinguish between visited states and success states*

1 User provides success examples



2 Unattended training



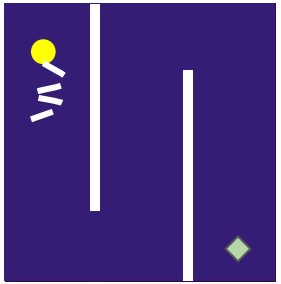
3 Successful policy

(Prior work: Fu et al., 2018; Singh et al., 2019, Zhu et al., 2020)

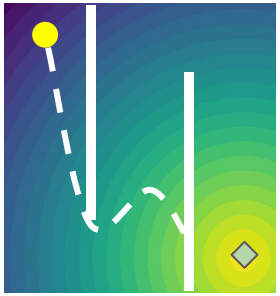
Toy Example: 2D Maze

Possible rewards

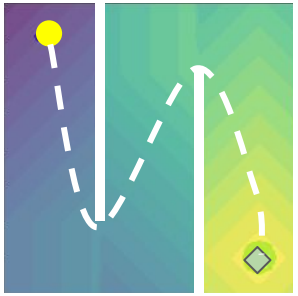
Sparse Reward
No reward signal



L2 Distance Reward
Misleading local optima

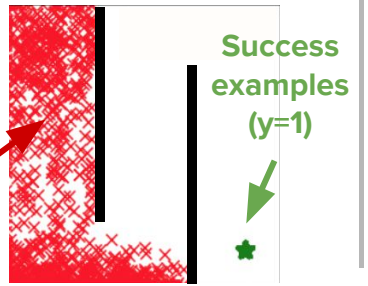


Ideal Shaped Reward
Guides agent toward goal

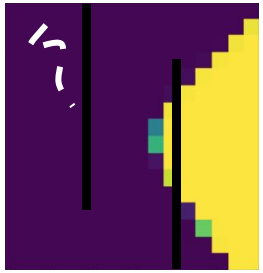


Standard classifiers

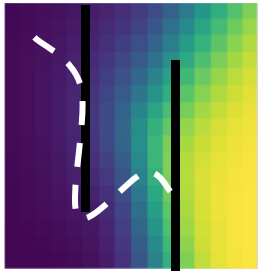
Training Set



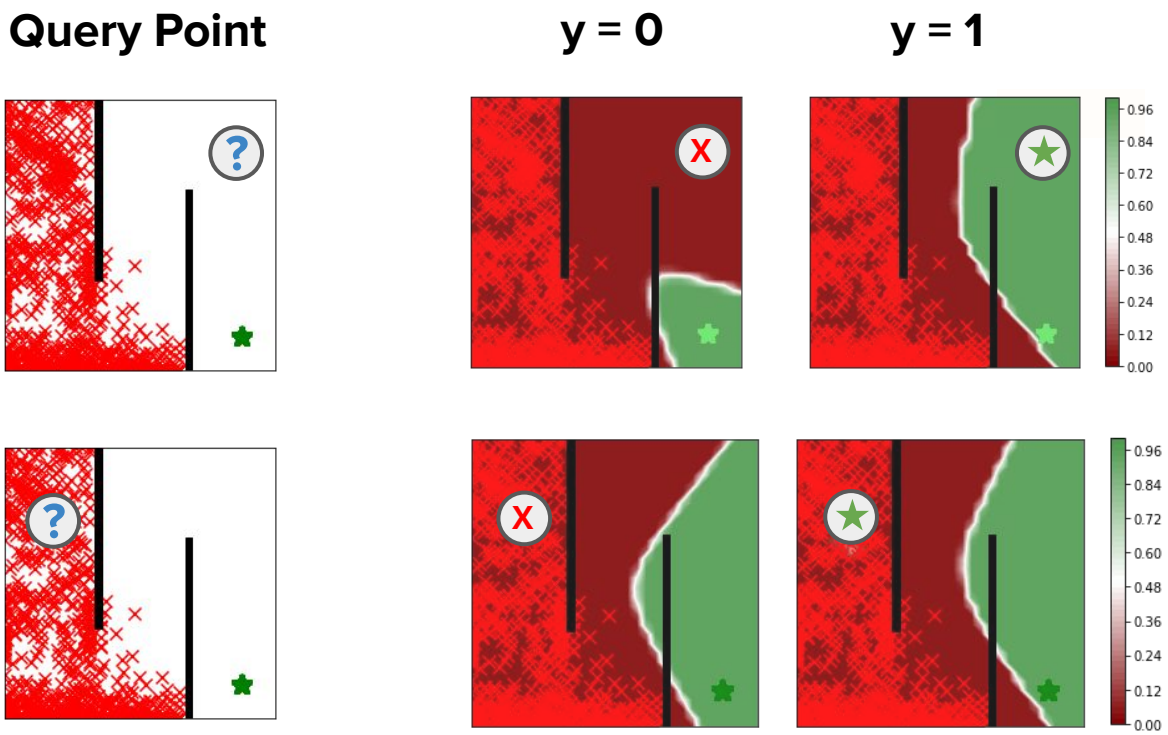
Trained to convergence
Little reward signal



Early stopping + L2 reg
Misleading local optima



Key idea: use **Normalized Maximum Likelihood (NML)** to obtain reward classifiers that are uncertainty-aware and provide better shaping.



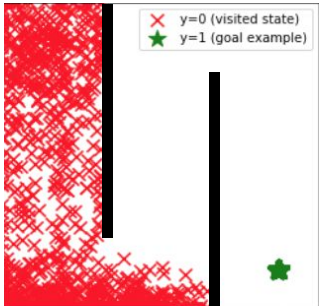
Key idea: use **Normalized Maximum Likelihood (NML)** to obtain reward classifiers that are uncertainty-aware and provide better shaping.

$$\theta_i = \arg\max_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim \mathcal{D} \cup (x_q, y=i)} [\log p_\theta(y|x)]$$

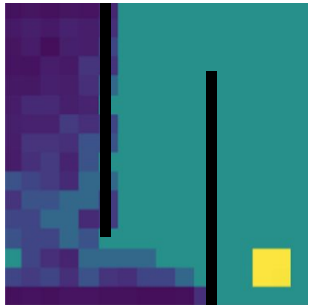
SLOW!!

$$p_{\text{CNML}}(y = i|x_q) = \frac{p_{\theta_i}(y = i|x_q)}{\sum_{j=1}^{k} p_{\theta_j}(y = j|x_q)}$$

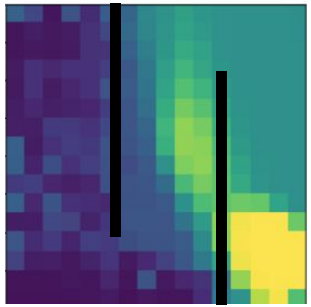
Training Set



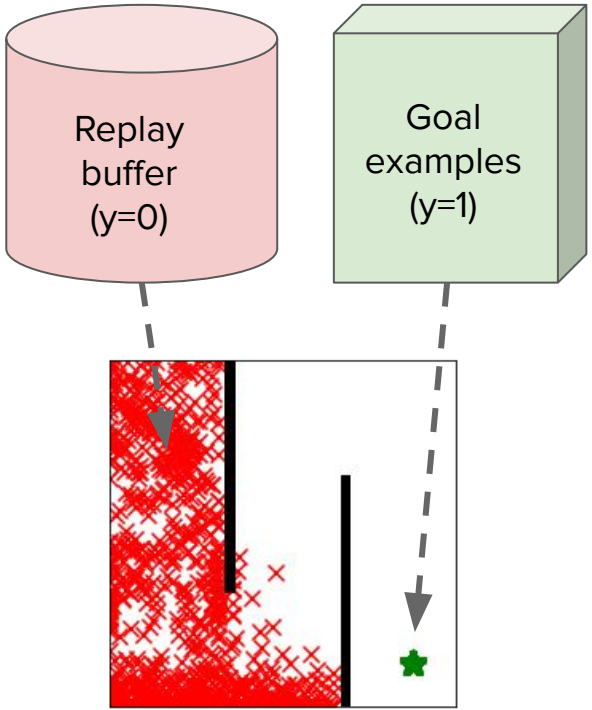
NML (discrete)



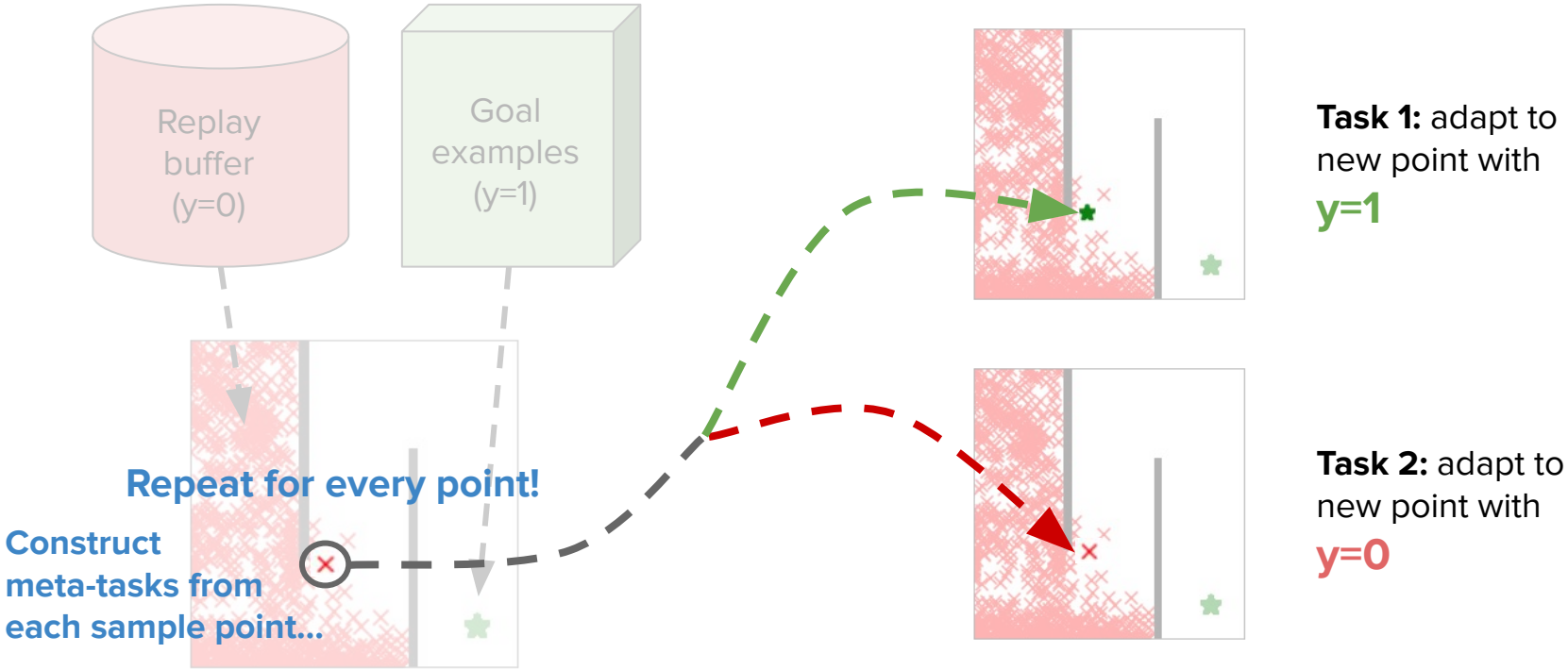
NML (with deep NNs)



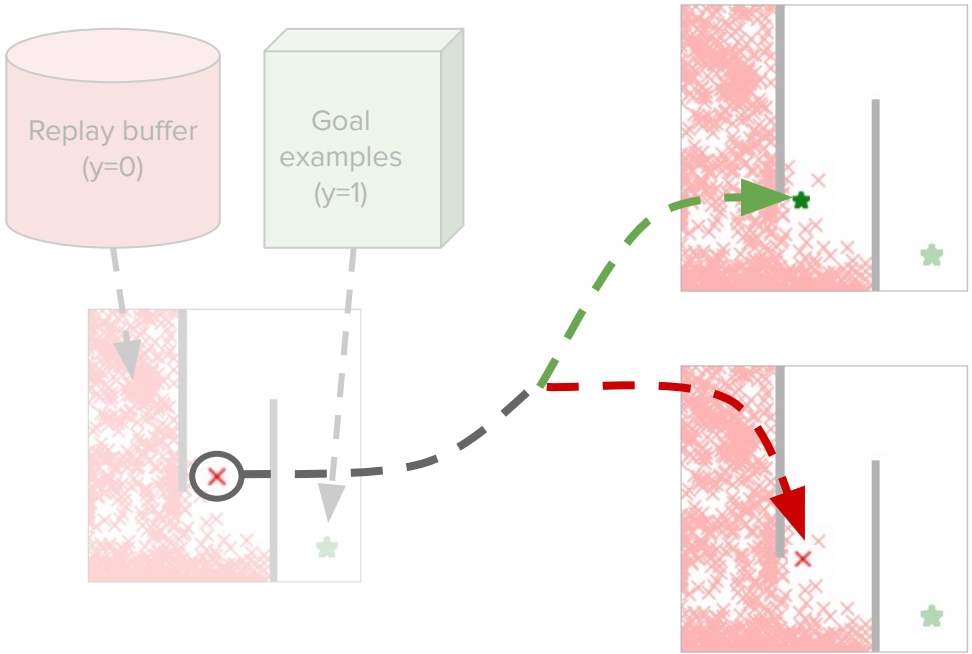
Meta-NML: efficient NML via meta-learning



Meta-NML: efficient NML via meta-learning



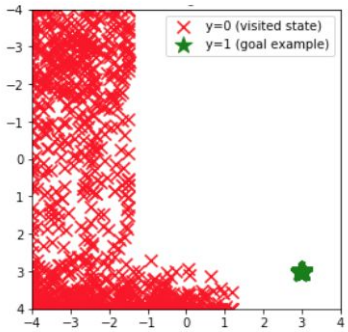
Meta-NML: efficient NML via meta-learning



Task 1: adapt to new point with **y=1**

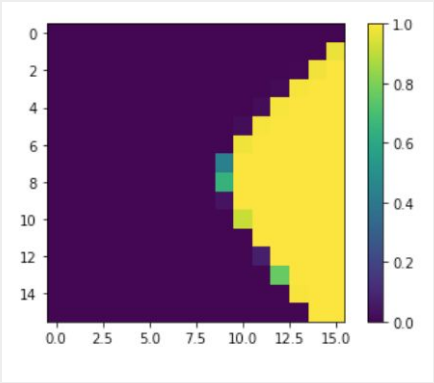
Task 2: adapt to new point with **y=0**

Training set



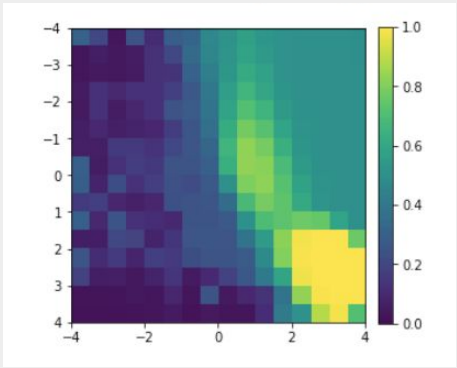
Meta-training

Model Initialization

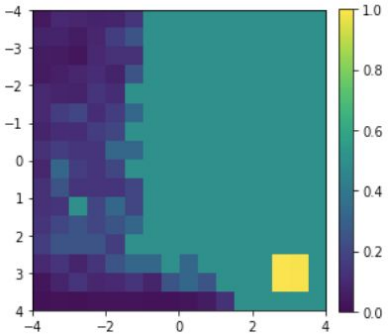


1 gradient step...

Meta-NML Rewards



(Ideal NML Rewards)

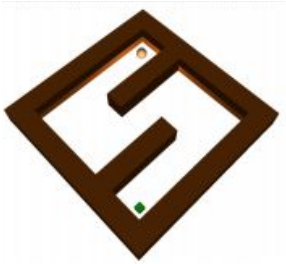


Runtimes for a single epoch of RL

MLE Classifier	Meta-NML	Naive CNML
23.50s	39.05s	4hr 13min 34s

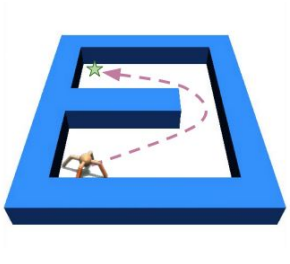
Evaluation Domains

Navigation Tasks



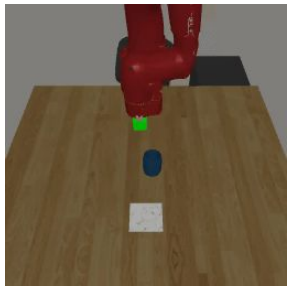
Zigzag Maze

Spiral Maze

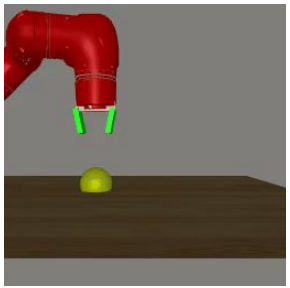


Ant Locomotion

Robotic Manipulation Tasks

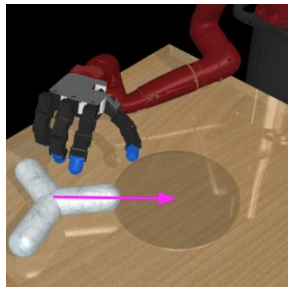


2D Pusher

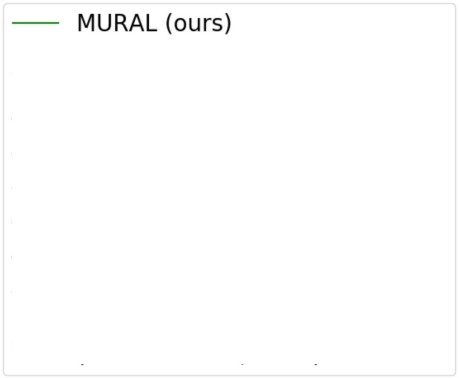


3D Pick-and-Place

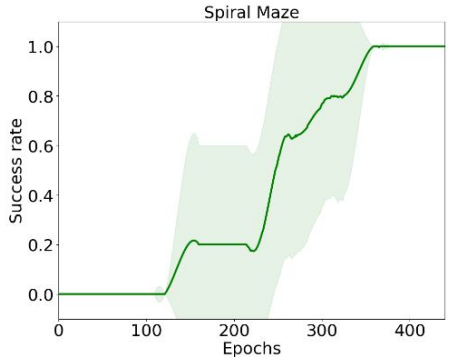
Door Opening

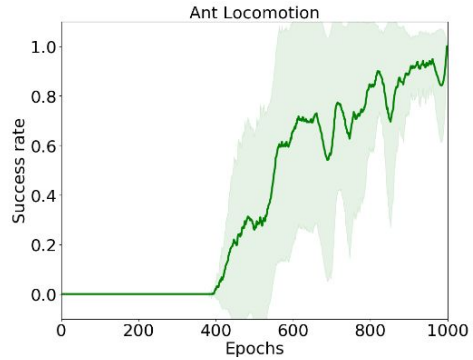


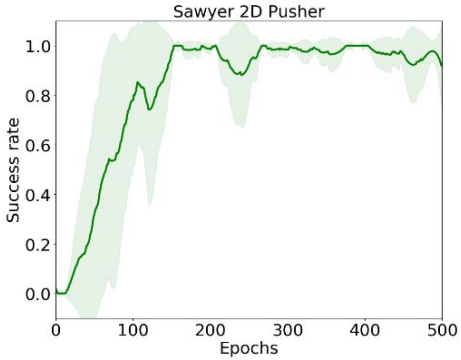
Dexterous Hand

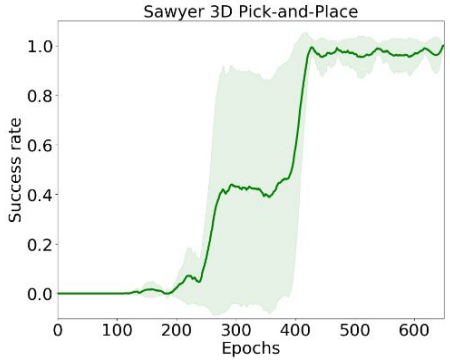

MURAL (ours)

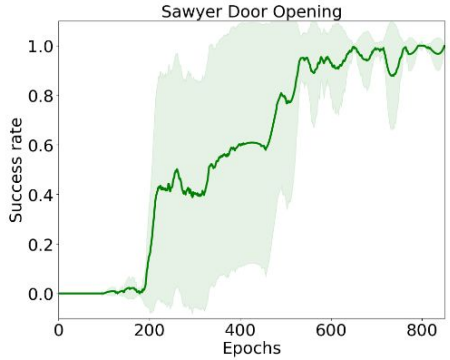

Zigzag Maze
Success rate
Epochs

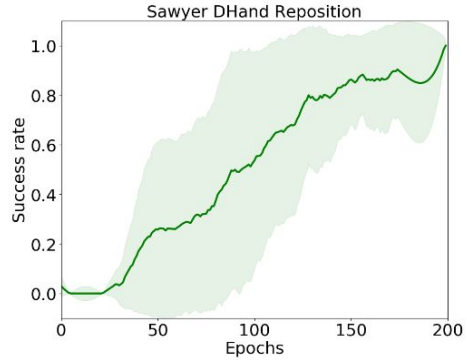

Spiral Maze
Success rate
Epochs

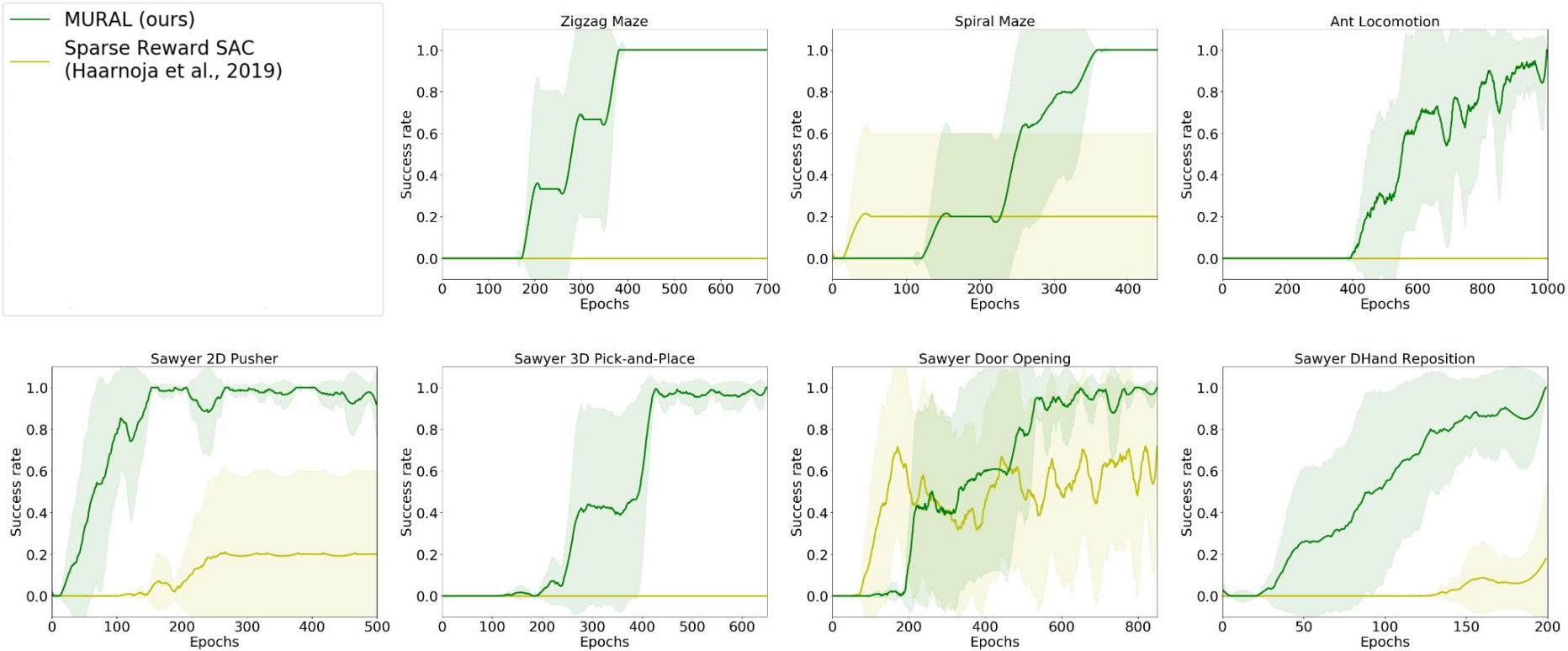

Ant Locomotion
Success rate
Epochs

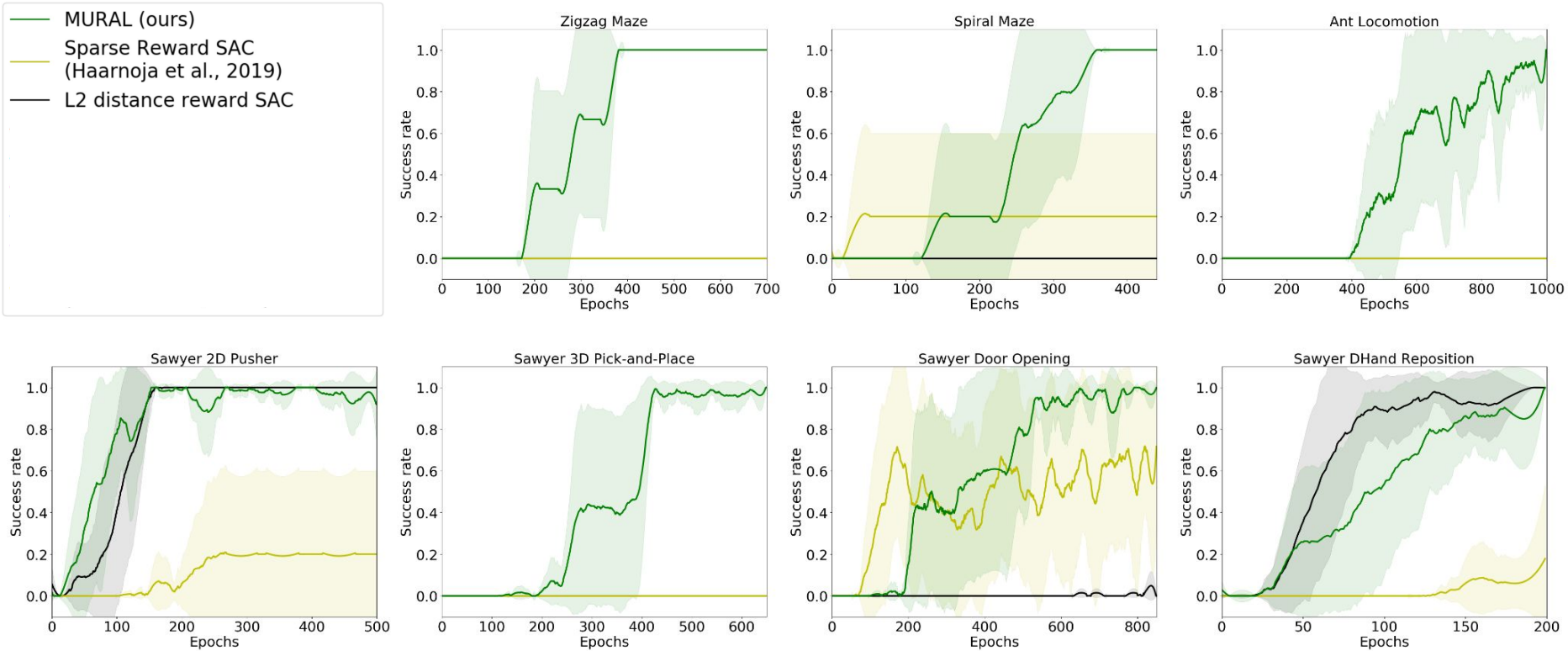

Sawyer 2D Pusher
Success rate
Epochs

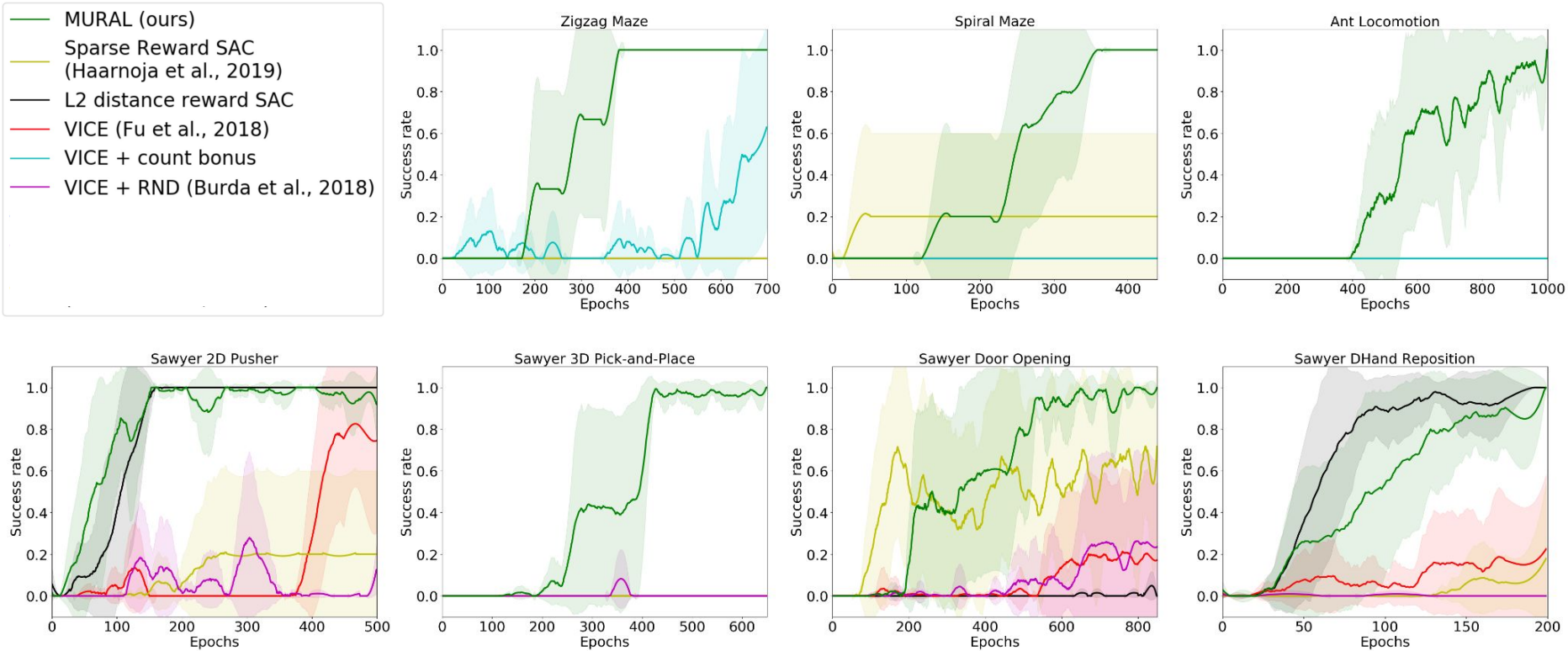

Sawyer 3D Pick-and-Place
Success rate
Epochs

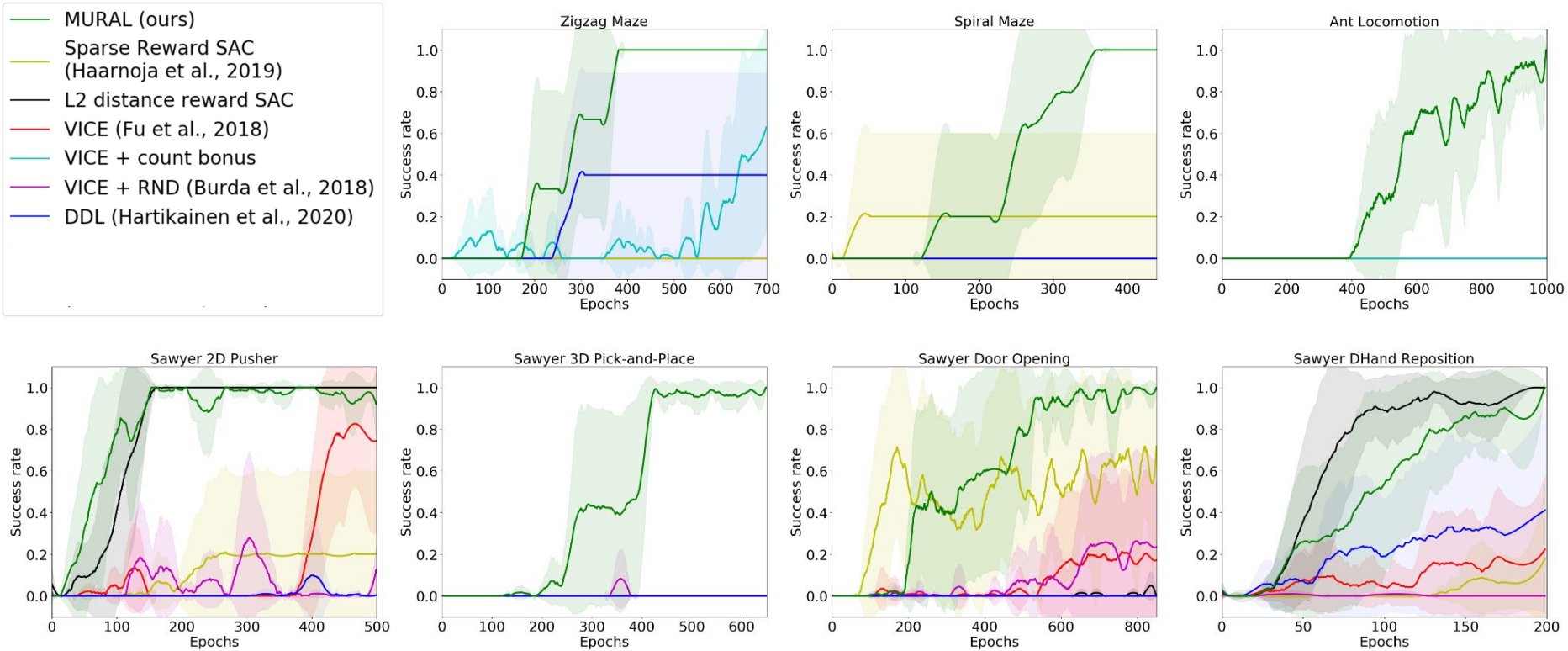

Sawyer Door Opening
Success rate
Epochs

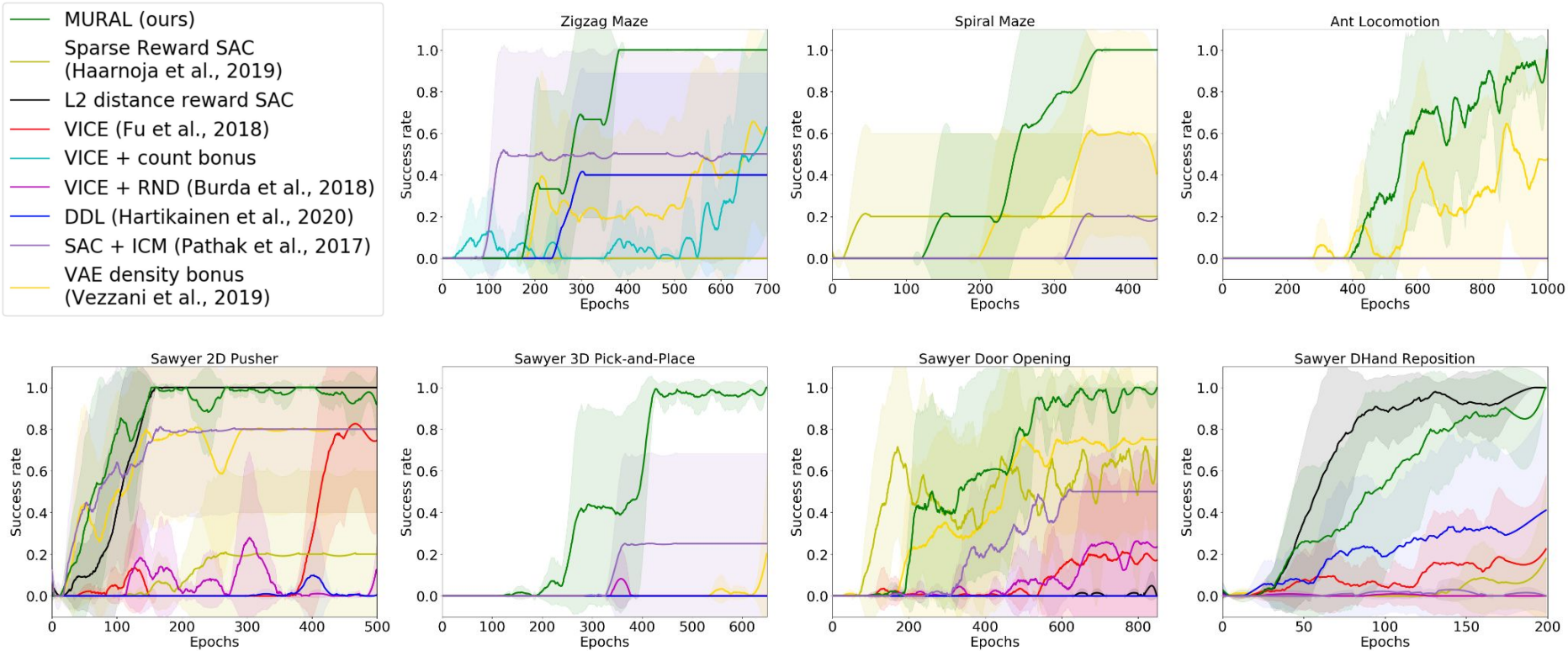

Sawyer DHand Reposition
Success rate
Epochs


MURAL (ours)
Sparse Reward SAC
(Haarnoja et al., 2019)
Zigzag Maze
Spiral Maze
Ant Locomotion
Sawyer 2D Pusher
Sawyer 3D Pick-and-Place
Sawyer Door Opening
Sawyer DHand Reposition
Success rate
Epochs


MURAL (ours)
Sparse Reward SAC (Haarnoja et al., 2019)
L2 distance reward SAC
Zigzag Maze
Spiral Maze
Ant Locomotion
Sawyer 2D Pusher
Sawyer 3D Pick-and-Place
Sawyer Door Opening
Sawyer DHand Reposition
Success rate
Epochs


MURAL (ours)
Sparse Reward SAC (Haarnoja et al., 2019)
L2 distance reward SAC
VICE (Fu et al., 2018)
VICE + count bonus
VICE + RND (Burda et al., 2018)
Zigzag Maze
Spiral Maze
Ant Locomotion
Sawyer 2D Pusher
Sawyer 3D Pick-and-Place
Sawyer Door Opening
Sawyer DHand Reposition
Success rate
Epochs


MURAL (ours)
Sparse Reward SAC (Haarnoja et al., 2019)
L2 distance reward SAC
VICE (Fu et al., 2018)
VICE + count bonus
VICE + RND (Burda et al., 2018)
DDL (Hartikainen et al., 2020)
Zigzag Maze
Spiral Maze
Ant Locomotion
Sawyer 2D Pusher
Sawyer 3D Pick-and-Place
Sawyer Door Opening
Sawyer DHand Reposition
Success rate
Epochs


MURAL (ours)
Sparse Reward SAC (Haarnoja et al., 2019)
L2 distance reward SAC
VICE (Fu et al., 2018)
VICE + count bonus
VICE + RND (Burda et al., 2018)
DDL (Hartikainen et al., 2020)
SAC + ICM (Pathak et al., 2017)
VAE density bonus (Vezzani et al., 2019)
Zigzag Maze
Spiral Maze
Ant Locomotion
Sawyer 2D Pusher
Sawyer 3D Pick-and-Place
Sawyer Door Opening
Sawyer DHand Reposition
Success rate
Epochs

MURAL: Meta-Learning Uncertainty-Aware Rewards for Outcome-Driven Reinforcement Learning

Kevin Li*, Abhishek Gupta*, Ashwin Reddy, Vitchyr Pong, Aurick Zhou, Justin Yu, Sergey Levine

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Website: https://sites.google.com/view/mural-rl